LCA Methodology

Data Quality and Uncertainty in LCI

Remi Coulon, Vincent Camobreco, Helene Teulon, Jacques Besnainou

Ecobalance, Inc.

Corresponding author: Remi Coulon, Ecobalance, Inc., 15204 Omega Drive, Suite 220, Rockville, Maryland 20850, USA

Abstract

It has recently been acknowledged that the quality of data used in Life Cycle Assessment (LCA) is one of the most important limiting factors to the application of the methodology. Early approaches dealing with this problem solely based on Data Quality Indicators (DQI) have revealed their limitations, and stochastic models are increasingly proposed as an alternative. Although facing methodological and practical difficulties, for instance the characterization of the distribution of input data, these stochastic models can significantly enhance decision-making in LCA. Uncertainty and data quality, however, are two distinct attributes. No matter how sophisticated the stochastic models are, they do not address the issue of the adequacy of the data used with regard to the goal of the study. Actual data on the distribution of SO_x emissions for US coal fired power plants for instance, would be of low quality for a European study. It is therefore believed that mixed approaches DQI/stochastic models should be developed in the future.

Keywords: Data Quality Indicators (DQI); data quality, LCA; decision-making in LCA; DQI, stochastic models; LCA; Life Cycle Assessment; quality of data, LCA; stochastic models; uncertainty, data quality

1 Introduction

In Life Cycle Assessment (LCA), as in any modeling approach, the quality of the data is one of the most important limits to the application of the methodology. The variability of the measurements within industrial plants, the discrepancy between bibliographical and actual site data and the sensitivity of the results to core methodological choices constitute, at different scales, obstacles to the reliability and comparability of LCA results. The influence of methodological choices such as co-product allocation and system boundaries can be readily assessed through the use of sensitivity analyses. On the other hand, the influence of data quality on the end results has rarely been analyzed. This paper presents some approaches that can be used in assessing the influence of data quality upon the final results and how the introduction of uncertainty in the data acquisition and analysis can significantly enhance decision-making in LCA.

Compared to the influence of methodological choices, the influence of the quality of data used upon final results has recently been acknowledged in the development of the LCA technique. Two routes have been taken in order to alleviate this concern:

- Data Quality Indicator (DQI): attributing a distinct indicator, separated from the input it qualifies. As seen in section 2, such indicators can encompass a variety of attributes such as completeness, reproducibility, etc.
- Use of stochastic models through the assessment and propagation of the inherent variability of the data itself (range, standard deviation, full distribution of the input, etc.). Section 3 details such approaches in terms of data acquisition and presentation.

2 Data Quality Indicator Approaches

The early approaches that were developed for instance by SETAC [1] and U.S.-EPA [2] to alleviate data quality concerns were based on an increasing number of data quality indicators. In SETAC [1], the list of relevant indicators is quite extensive, including indicators that can be both:

- **Quantitative:** Accuracy, bias, completeness, distribution, homogeneity, precision and uncertainty
- **Qualitative:** Accessibility, applicability, comparability, consistency, derived models, identification of anomalies, peer review, representativeness, reproducibility, stability and transparency.

Applying this rigorous data documentation system to every piece of data throughout an actual size LCA (e.g., including several hundreds of processes) is a time consuming exercise

Int. J. LCA 2 (3) 178 - 182 (1997)

which can become almost impractical for large LCA projects. The presence of this detailed information, however, is quite useful when it comes to tracking down the source and quality of a particular data point.

Nevertheless, the biggest objection to these systems is that they do not provide a summary statement of the data quality of a final LCA result, such as total CO_2 emissions for the life cycle of an appliance. Although the individual CO_2 emissions for each and every step of the system might be precisely characterized with a complex system of data quality indicators (also termed data pedigree), no condensed index is provided for the total, nor are any means of conveying the uncertainty of the results.

The following example is based on the simple cradle-to-gate system detailing the production of a hammer. As carried out by the hammer manufacturer, the system could be represented as in **Figure 1** showing in great detail the steps that the manufacturer controls (assembly steps) while condensing the suppliers' processes into one black box. Let us further assume that the data quality of each process step is characterized by a given DQI (such as completeness or representativeness), represented by a letter. It can be imagined that the scoring system ranges from A (best data) to E (worst data).

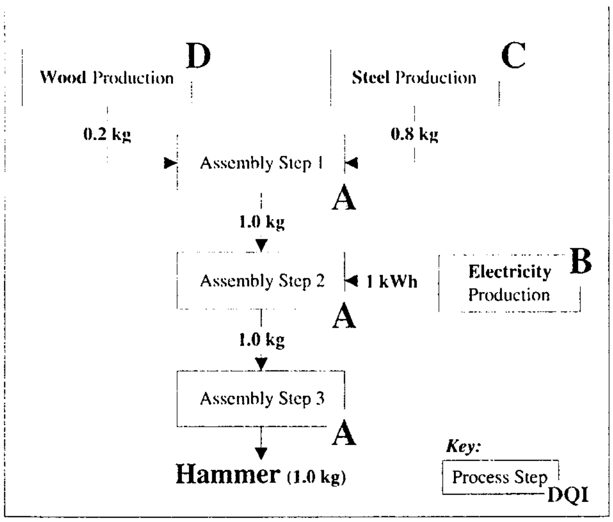

Fig. 1: Determination of an overall data quality indicator. Example system

As can be seen in the above example, no overall DQI can be easily computed for this system without significant value judgment. In an attempt to build a simple average of the DQIs, one will end up with a disproportionate influence of the assembly steps (which are likely to entail less environmental impacts than the upstream processes) on the overall system. Using a weighted average will not resolve the issue since the assembly steps will still carry an excessive weight on the overall score. The structure of the system, which is likely to vary among different users' characterization of the same industrial system, substantially influences any such overall DQI. Moreover, it is difficult to account for all electricity producing processes in a weighted average approach since there is no correspondence between mass end electricity.

Therefore, although it is important to keep track of detailed data assumptions, these approaches, as currently promoted under ISO [3], provide what has been termed a "post-it note" approach to data quality where each process step may be accurately described but no aggregate information can be produced reliably.

Even when coupled with sensitivity analyses, these approaches do not provide a consistent framework for analyzing the variability and uncertainty of the final results. Sensitivity analyses can indicate the extreme values that a system can take, but they do not provide any information on the distribution within that range (i.e. do we attain the 5% lower values in 5% or 30% of the cases?) It is this type of shortcomings that have motivated the development of stochastic models.

3 Stochastic Model Approaches

Only more recently were approaches developed to build summary indicators accounting for uncertainty in a quantitative fashion. Such approaches have been introduced several years ago in risk assessment and are rapidly gaining ground. U.S.-EPA's Science Policy Council recently approved a new agency policy on probabilistic risk assessment using Monte Carlo and other analytical methods, opening the door to significant increases in the use of such approaches and to more complete analysis of uncertainty and variability (Risk Policy Report [4]).

The mathematical framework for handling and propagating uncertain quantities is well known, and the method of choice is the Monte Carlo method, as used in Kennedy [5] and Besnainou [6].

The Monte Carlo method simply re-calculates the final results a large number of times, with each calculation making use of a value for each uncertain parameter which is drawn at random from the specified input distribution.

Thus, a value for each input parameter is drawn and the final result computed. After repeating this process several hundreds of times, the distribution of the final results can be plotted, providing a reasonably stable approximation for the shape of the output variable's distribution which would have to be obtained from an infinite number of such trials.

Therefore, the issue for LCA relies more on the actual determination and characterization of the uncertainty of input data.

3.1 Acquisition of probabilistic data

There are two possibilities for acquiring information on the distribution of input data in an LCA:

- Actual data
- Expert judgment (KENNEDY [5] and WEIDEMA [7]).

Actual data evidently provides the best source of information for characterizing the distribution of the input. Figure 2 indicates the probability density function of SO_2 emission factor in grams per ton from a US-EPA database on utility coal boilers.

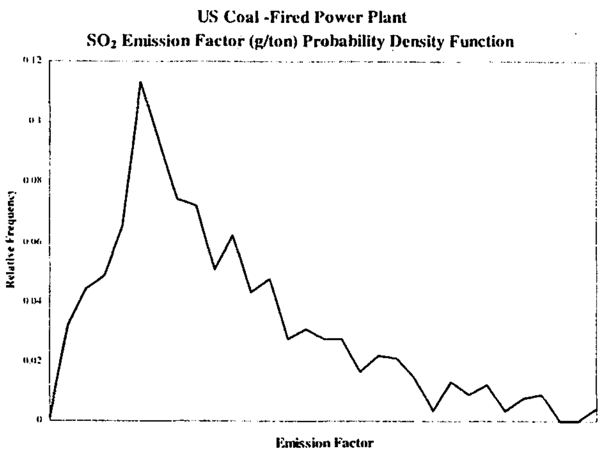

Fig. 2: U.S. coal-fired power plant. SO_2 emission factor (g/ton) probability density function

It is interesting to notice that not all LCI data categories present the same uncertainty. For instance, unpublished data from the Association of Plastic Manufacturers in Europe were used to estimate the ranges of values for plastic materials. Energy consumption figures are relatively comparable among different producers and do not vary much with time since energy is closely related to the process efficiency which depends more upon the basic thermodynamics of the process than on its actual operation (the latter are also minimized for reasons of costs). Air emissions, water effluents and solid waste on the other hand, vary quite considerably among producers, reflecting differences in processes and especially the various regulations they must submit to. Table 1 presents the ranges that were found for various inventory data categories. In all cases, the average has arbitrarily been set to one.

However, because inventory flows are interrelated, the notion of best and worst sites should be interpreted carefully, and it is highly improbable that one site will have the minimum values (or maximum) for all LCI data categories. Let us suppose that sites A and B, having the same process yield and gross waste scrap generation, are among the sites included in the average of Table 1.

Table 1: Range of the various inventory data categories. Polymer production

	Min.	Average	Max.
Energy	0.86	1.0	1.3
Main air emissions	0.18	1.0	2.9
Water effluents	0.01	1.0	17.0
Hazardous solid waste	0.00	1.0	21.0
Non-haz. solid waste	0.05	1.0	2.8

However, since site A implements an aggressive solid waste reduction plan, most of its waste is treated (either on-site or off-site), yielding a higher fossil fuel consumption and therefore higher air emissions than site B (→ *Fig. 3*). Because LCA results between different media are often interrelated, and not always in the same manner (i.e. if A generates less waste than B, it does not imply that A generates less air emissions than B), care has to be taken when speaking about "worst" or "best" sites in such averages. In Figure 3, for example, site A is simultaneously best for waste generation and worst for air emissions and fossil fuel consumption.

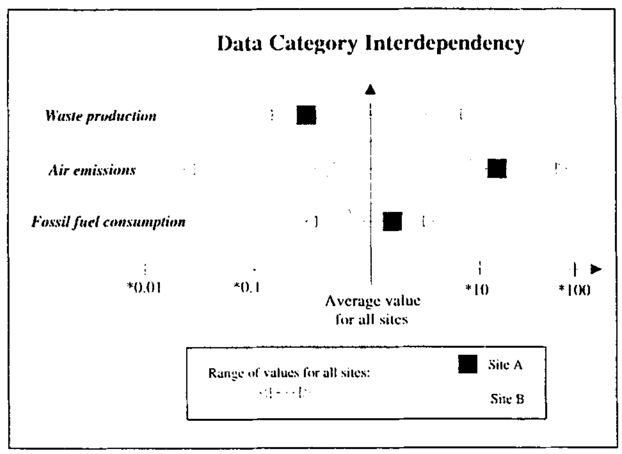

Fig. 3: Data category interdependency

When no actual data is available, expert judgment has to be used to quantify the distribution of the input data. We have been using mostly lognormal, triangular and uniform distributions. Lognormal and triangular distributions (symmetric or not) are used in most cases, and are appropriate when there is strong reason to expect a central tendency to the data. Uniform distributions are used when strong central tendencies are not expected.

The use of expert judgment in the characterization of inputs leads to a wider garbage-in, garbage-out problem, and the qualitative determination of data uncertainties can arguably be subject to criticism. Practical approaches based on simplified pedigree matrix used for quantifying uncertainties should be explored, as in KENNEDY [5] and WEIDEMA [7].

These pedigree-based approaches provide practitioners with a consistent framework to formulate assumptions, but are not exempt from criticism either.

3.2 Result presentation

Results are presented as illustrated in **Figure 4**, indicating the minimum and maximum values as well as the 5th percentile (value in the data set for which 5% of the values are below it) and 95th percentile (value in the data set for which 5% of the values are above it). The 90% confidence interval (the probability that the value is contained within the interval is 90%). The indicated results are extracted from a study carried out for a chemical company on the cradle-to-gate comparison of Polyvinyl Chloride (PVC) and Polycarbonate/ Acrylonitrile Butadiene Styrene (PC/ABS) for use in durable goods. The compared indicator is the Acidification Potential, expressed in gram equivalent H+ (and based on inventory flows such as SO_x, NO_x).

Figure 4 indicates large overlaps and does not allow for the clear identification of a better option. Note that the selection of the percentiles is arbitrary (a 95% or 99% confidence interval could have been selected) and is a function of the willingness of the decision-maker to take risks. Also up to the decision maker is the importance accorded to minimum and maximum values. Whereas a risk assessment approach could be more concerned with extreme cases, for instance, an LCA approach could provide more emphasis to the "bulk" of the data rather than to extreme values.

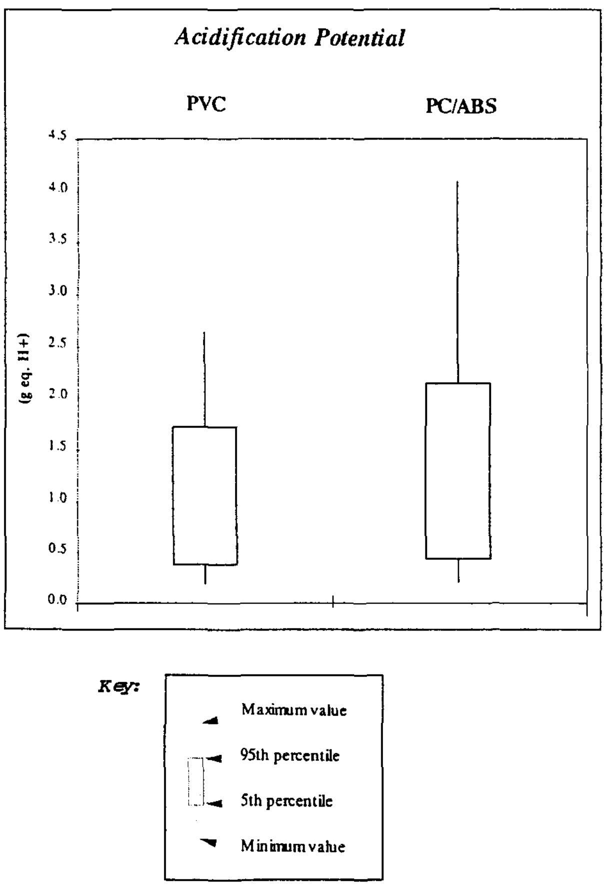

Fig. 4: Example of results using uncertainty analysis

The two studied materials rely on a number of similar processes (such as distillation of crude oil, chlorine production for PVC and phosgene) for which the same data sets have been used. Therefore, the previous results do not convey the potential correlation between the two options. For instance, the acidification potential shows significant ranges for both options with potentially large overlaps, although these two quantities might be strongly correlated, indicating that the difference between the two options might not be as large as suggested.

A way to eliminate the effect of such a correlation is to study the probability distribution of their normalized difference. Therefore, the same comparison of PC and PVC can be analyzed in terms of the following ratio (Norris [8]):

$$\frac{(PVC - PC/ABS)}{PVC}$$

This ratio, therefore, indicates how PC/ABS compares to PVC:

- A 10% figure indicates that PC/ABS is 10% better,
- a -60% figure indicates that PC/ABS is 60% worse.

Figure 5 shows how such a ratio can be interpreted from a stochastic point of view.

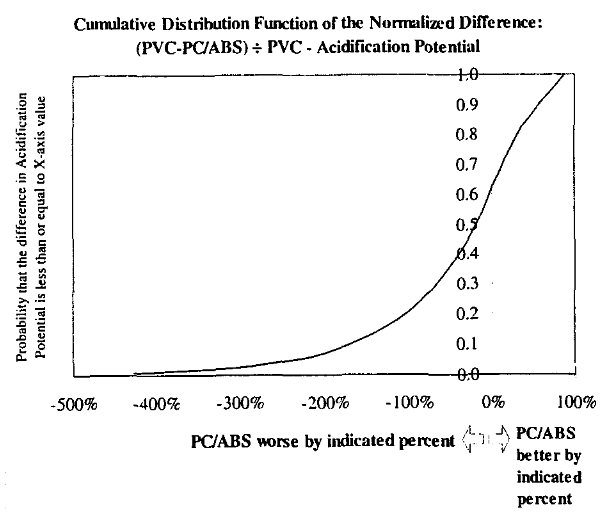

Fig. 5: Acidification potential for the PVC vs. PC/ABS comparison. Cumulative distribution function of the normalized difference between the two options

If 10% is considered a significant difference between the two options it can then be said that there is a 28% chance that PC/ABS is better (by 10% or more) and a 58% chance that PC/ABS is worse (by 10% or more) than PVC.

This type of result can be worthwhile in decision-making and helps to distinguish more clearly between the two options in our example on Acidification Potential. However, this approach also adds a layer of complexity to the readability of the results. In cases where the compared alternatives rely on markedly different technologies and share little if any common processes, this additional step can be omitted.

4 Uncertainty vs. Data Quality

However useful stochastic models are in terms of conveying the uncertainty of LCA results to decision makers, they fall short of addressing the true quality of the data, or rather the adequacy of the data used, with regard to the scope of the project.

Uncertainty and data quality are two different attributes since data on which all statistical information could be available (including the full distribution of the value) may yet be of low quality for its function. Actual data on the distribution of SO_x emissions for US coal fired power plant (for data which exists on the distribution, see *Fig. 2*), for instance, would be of low quality for a European study.

This distinction can be related to the distinction made by the British Standards for Quality Assurance between grade and quality (FUNTOWICZ and RAVETZ [9]). The grade denotes the general degree of refinement and elaboration of a product or a data set. The quality denotes instead the fitness for purpose of the data or product.

However sophisticated the stochastic models are, they do not address the issue of the data used with regard to the goal of the study.

Mixed approaches using Monte Carlo/DQI (as in WEIDEMA [7]) are being explored, but since, as seen in Section 2, DQI cannot be propagated for an overall system, this implies that DQI should modify the distribution themselves before they are propagated. This adjustment of distribution curves due to the use of not adequate data (from a geographical, temporal and technological point of view) can entail significant value judgment. Here, it is always recommended that broad ranges be used in order to be conservative.

5 Conclusion

After having carried out several LCA projects using stochastic approaches, our main conclusions are that:

- The use of probability distribution in the characterization of inputs increases the duration of the data collection phase. First order estimates can be derived quickly from professional judgment, past experience and similar studies, but their validation requires time. Once such probabilistic databases are built however, they can be used directly in other projects.
- The use of stochastic models or combinations of DQI and stochastic models is still a research field, and its application should be reserved to selected case studies. Particularly, the nature and influence of data category interdependency should be explored in greater detail.
- The use of stochastic models and the presentation of ranges and confidence intervals enhances decision-making by helping to focus on categories for which there are real differences between alternatives.
- The benefit of presenting such data ranges to decision-making audiences compensates the potential pitfalls of such probabilistic methods. It is believed that explicitly presenting ranges of values instead of "magic numbers" can greatly reduce the potential of intentional or unintentional misuse of the LCA technique. Moreover, with audiences increasingly aware of the complexity and uncertainty underlying the final results, LCA should provide estimates about ranges or levels of confidence in the results.
- Since the LCA stochastic model will tentatively account for and present distribution of the endpoints, their results will carry more weight than those of deterministic LCA models. Therefore, such stochastic approaches put an even greater emphasis on the need of transparency in the communication of the results as well as the intensive use of sensitivity analysis. High quality does not require the elimination of uncertainty but rather its effective management and presentation.

6 References

[1] Life Cycle Assessment Data Quality: A Conceptual Framework. 1994. Society of Environmental Toxicology and Chemistry. 1994. Pensacola, Florida

[2] Guidelines for Assessing the Quality of Life Cycle Inventory Data. 1995. U.S. Environmental Protection Agency. EPA/ 530-R-95-010

[3] Environmental Management - Life Cycle Assessment - Goal and Scope Definition and Inventory Analysis. March 5, 1997. Draft International Standard 14041. International Organization for Standardization

[4] RISK POLICY REPORT - February 21, 1997, Vol. 3 N. 2 p 6

[5] KENNEDY, D.J. et al: Stochastic Environmental Life Cycle Assessment Modeling. 1996. Int. J. LCA 1 (4) 199-207

[6] BESNAINOU, J. and COULON, R.: Uncertainty, Complexity and Decision in LCA. Proceedings of the second international conference on EcoBalance. Nov. 18-20, 1996. Tsukuba, Japan

[7] WEIDEMA, B.P. and WESNOES, M.S.: Data Quality Management for Life Cycle Inventories - An Example of Using Data Quality Indicators. Presented at the 1995 SETAC World Congress

[8] NORRIS, G.A.: Picking Winners and Telling Alternatives Apart: Practicing LCIA Decision Making in the Presence of Uncertainty. June 1996. 89th Annual Meeting Air & Waste Management Association

[9] FUNTOWICZ, S. O. and RAVETZ, J.R.: Uncertainty and Quality in Science for Policy. Kluwer Academic Publishers, The Netherlands